Concordanciers et flexion automatique

Eric Laporte

Concordanciers et flexion automatique

ÉRIC LAPORTE
Université Paris-Est Marne-la-Vallée - IGM-Labinfo

Introduction

Des linguistes de plus en plus nombreux utilisent des concordanciers pour extraire de textes préexistants des mots ou expressions en contexte. L'objectif de cet article est d'examiner plus particulièrement ceux qui produisent des concordances lemmatisées, c'est-à-dire qui regroupent les différentes formes fléchies d'un même mot. Cette caractéristique technique est liée à un traitement sur les dictionnaires : la flexion automatique. Nous expliquons en quoi consiste cette opération et ce qu'elle implique pour le lexicologue.

La première partie de cet article passe en revue les utilisations des concordances. Nous nous intéressons ensuite aux principaux critères de qualité qui permettent de choisir un concordancier satisfaisant pour un usage donné. La troisième partie concerne les concordanciers avec dictionnaire, capables de produire des concordances lemmatisées. La quatrième partie est consacrée à la flexion automatique, un traitement de dictionnaires indispensable à la maintenance, et parfois au fonctionnement, d'un concordancier de ce type.

1. À quoi servent les concordances ?

Les concordances sont un outil de base du linguiste cherchant à étudier les contextes d'emploi de mots ou d'expressions dans des textes préexistants. Un corpus de textes étant fixé, une concordance est la liste de toutes les occurrences d'un ou plusieurs mots ou expressions, alignées verticalement en colonne, accompagnées de leur contexte droit et gauche (fig. 1), et souvent classées, par exemple par ordre alphabétique.

```
   Mais un nouveau tour de vis sera nécessaire pour ramener le point d'équilibre au-des
ational. « Quels hommes seront donc nécessaires pour l'industrie de demain ? Comment
disposera du « plus » de croissance nécessaire. Il est alors prévu de rendre cette
tenues de leur fournir l'assistance nécessaire - rechercher tous les éléments d'informa
on parvenir au degré de convergence nécessaire, dans un marché unique très intégré, tou
   s. Mais ce souci de transparence indispensable ne doit pas pour autant peser sur la
sûr les masques tombent - référence obligée à la fête au château de la Règle du jeu, de
e que « le FLN représente une force nécessaire pour l'équilibre politique de la société
   e dynamique. Et le coup de pouce indispensable qui permet bien souvent d'emporter un
```

Figure 1. Extrait d'une concordance.

Un concordancier est un outil informatique qui produit des concordances à partir de corpus de textes et de requêtes. L'utilisation de concordanciers s'est développée au cours du dernier quart du XX^{e} siècle avec la vogue des corpus en linguistique. Elle correspond à un progrès réel car elle facilite la rigueur dans l'observation des faits. Les concordanciers peuvent servir de nombreux buts théoriques et pratiques. Nous allons en évoquer quelques-uns.

Des concordances produites à partir de textes de qualité facilitent la construction de dictionnaires pour les lecteurs humains. Le dictionnaire *Cobuild* de l'anglais (J. SINCLAIR, 1987) en est l'exemple prototypique. Les exemples qui font sa richesse sont tirés d'un corpus de centaines de millions de mots de textes de qualité. À l'époque, la production automatique de concordances à une telle échelle avait nécessité des développements informatiques spécifiques.

Aux dictionnaires pour les lecteurs humains, ou dictionnaires conventionnels, nous opposons ceux pour le traitement des langues, ou dictionnaires électroniques. Les différences entre ces deux types de dictionnaires, et de lexicographie, sont profondes : les définitions et exemples sont propres aux dictionnaires conventionnels, alors que les exigences de codification et d'exhaustivité sont propres aux dictionnaires électroniques (M. GROSS, 1989). Les concordances sont utilisées pour l'extension des dictionnaires électroniques, et notamment des dictionnaires de mots composés (M. SILBERZTEIN, 1993 : 183). Ainsi, une concordance telle que celle de la fig. 2, produite à partir d'une requête décrivant la séquence *N de N*, où *N* désigne la catégorie grammaticale substantif, permet au lexicologue de détecter des mots composés à coder dans les dictionnaires.

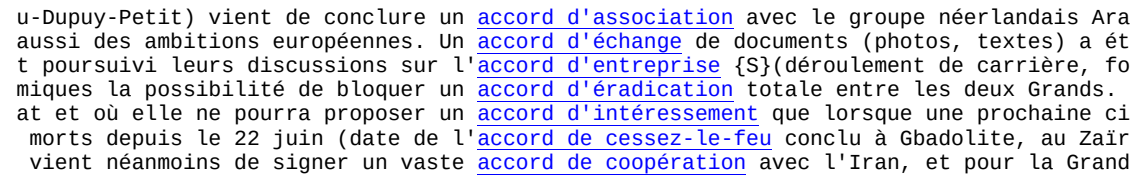

Figure 2. Extrait d'une concordance de séquences *N de N*.

Les philologues, qui manipulent des corpus de textes fermés et cherchent à en tirer toute la substantifique moelle, ont naturellement recours à des concordances. C'est d'ailleurs l'origine étymologique de cet emploi du mot *concordance*. Au XIII^{e} siècle, il s'agissait de listes d'occurrences d'un mot pour en illustrer les différents emplois, puis de concordances de tous les mots simples de la Bible en latin.

Dans les études stylistiques et littéraires, dans le domaine de l'édition critique, les concordances sont des outils appréciés quand les textes littéraires sont disponibles sur format électronique (M. R. CABALLERO, 1999; V. MAGRI-MOURGUES, 2006). Ils peuvent servir, par exemple, à étudier l'évolution d'un texte littéraire de version en version, quand celles-ci sont disponibles sous forme électronique.

Beaucoup de chercheurs en syntaxe et en sémantique utilisent systématiquement des concordances pour recueillir des exemples. Cette pratique tend à rendre plus rigoureuse la collecte des exemples, et à orienter l'étude vers les formes réellement en usage. Elle facilite aussi la construction de grammaires syntaxico-sémantiques. Ainsi, M. GROSS (2000) décrit une méthode générale de construction de grammaires locales[1] à l'aide d'un corpus de textes et d'un concordancier.

Pour les professionnels de la traduction (Ch. JACQUET-PFAU, 1994), des langages de spécialité, de la terminologie (H. ZINGLÉ, 1994 ; D. GOUADEC, 1997), les concordances sont le moyen le plus efficace d'explorer l'utilisation d'un mot, d'une expression ou d'un terme technique dans un type de texte donné ou dans un domaine technique.

L'enseignement des langues a recours à des concordanciers de plusieurs façons. On peut faire analyser des concordances par des apprenants pour leur présenter du vocabulaire, des phénomènes morpho-syntaxiques et les constructions syntaxiques propres aux éléments lexicaux. Une variante ludique de cet exercice consiste à présenter une concordance et faire deviner la requête, c'est-à-dire la description du motif linguistique que le concordancier a recherché dans le texte. Une autre utilisation consiste à extraire des phrases de textes préexistants pour créer ou adapter des exercices (Ch. TRIBBLE, G. JONES, 1997; M. GARRIGUES, 1998-1999; E. TOGNINI-BONELLI, 2001).

[1] Une grammaire locale est une représentation formalisée d'un ensemble d'expressions ou de séquences linguistiques d'une productivité limitée.

Pour certains de ces usages, notamment en lexicologie, en linguistique et dans l'enseignement des langues, la Toile peut être utilisée comme un corpus, un moteur de recherche jouant le rôle du concordancier. Comme les moteurs de recherche ne présentent généralement pas leurs résultats sous forme de concordances, des outils spécifiques ont été développés (A. RENOUF *et al.*, 2007).

Nous n'avons pas mentionné le cas où un texte est accompagné de sa traduction dans une autre langue, formant un « bitexte ». Cette situation ouvre des potentialités très prometteuses, et certains concordanciers s'y adaptent progressivement. Nous n'aborderons pas plus ici ce domaine passionnant, car il pose des problèmes spécifiques.

2. Critères de qualité d'un concordancier

Plusieurs concordanciers sont actuellement en usage et se distinguent par leurs conditions d'acquisition et leurs caractéristiques techniques. Il est naturel pour leurs utilisateurs de s'intéresser à la qualité de ces outils, car leur confort de travail et donc leur efficacité sont en jeu. Dans cette partie, nous examinons les principaux critères de qualité.

Deux paramètres, le rappel et la précision, sont particulièrement cruciaux car ils mesurent la capacité du concordancier à atteindre précisément la cible recherchée par l'utilisateur. Il s'agit des deux mêmes paramètres que ceux qui mesurent les performances des moteurs de recherche utilisés pour trouver sur la Toile les pages concernant un sujet donné. Le rappel mesure la capacité d'un moteur de recherche à trouver des pages correspondant au désir de l'utilisateur, par exemple à trouver des pages contenant *magistrats* et *justiciables* lorsque la requête comporte le mot *Justice*. Inversement, la précision mesure sa capacité à écarter les pages qui n'y correspondent pas, par exemple celles qui comportent la séquence *est rentrée à la maison* lorsque l'utilisateur visait la *rentrée* (des classes).

De même, un concordancier comporte une fonction de recherche, parfois appelée moteur de recherche, qui sert à détecter dans le corpus les occurrences auxquelles seront consacrées les lignes de la concordance. Si une requête vise l'expression *aller à la plage*, la difficulté pour le concordancier concerne le rappel, car elle consiste à détecter dans le texte les formes conjuguées très irrégulières telles que *ira* ou *va à la plage*. Si une requête vise les occurrences du point cardinal *est*, le défi concerne au contraire la précision, car le mot *est*, forme fléchie du verbe *être*, ne fait par partie de la cible. Le rappel et la précision dépendent certes de l'utilisateur, qui exprime ses requêtes avec plus ou moins de talent, mais aussi du concordancier. En effet, celui-ci doit, d'une part, permettre au lecteur d'exprimer des requêtes pertinentes en respectant les conventions du langage de requêtes du système, et d'autre part trouver ensuite précisément la ou les cibles correspondantes.

Le rappel et la précision étant au cœur du fonctionnement d'un concordancier, il vaut la peine de s'y arrêter. Or les requêtes font assez souvent et naturellement intervenir la notion de lemme, c'est-à-dire sont neutres par rapport à la flexion. Par exemple, un utilisateur intéressé par l'expression *aller à la plage* a des chances de conserver cet intérêt si les occurrences prennent les formes *ira* ou *va à la plage*. (Nous appelons lemme d'un mot celle de ses formes fléchies qui est choisie pour représenter toutes les autres, par exemple l'infinitif des verbes dans le cas du français. La notion de flexion regroupe celles de conjugaison, déclinaison, changement de genre grammatical, de nombre, etc.) Le rappel et la précision des concordances engendrées dans un tel cas dépendent fortement du système utilisé. La notion de lemme a une existence dans le langage de requêtes de certains systèmes, qui permettent d'exprimer une requête d'une façon neutre par rapport à la flexion d'un ou de plusieurs des mots qui la composent. Ainsi, <*aller*> *à la plage* pourra viser les séquences dans lesquelles

aller est conjugué de n'importe quelle façon possible, tout en se restreignant à celles où *plage* est au singulier. La concordance obtenue est parfois qualifiée de concordance lemmatisée.

D'autres systèmes n'offrent pas cette fonctionnalité, mais compensent par des requêtes dans lesquelles une partie des mots est libre, comme dans *jou* au football*, qui visera des séquences dont le premier mot commence par les trois lettres *jou* suivies d'un nombre indéterminé de lettres quelconques. Lorsque l'utilisateur exprime une telle requête, il est souvent amené à faire un compromis entre rappel et silence. Ainsi, *joue** ne reconnaîtra pas *jouait*, ce qui peut diminuer le rappel dans la concordance obtenue, car des séquences comme *jouait au football* ne seront pas détectées. Inversement, *jou** reconnaîtra aussi *jour*, ce qui peut diminuer la précision, car le système pourra présenter à l'utilisateur *revenir un jour au football traditionnel*. De telles requêtes constituent une approximation de la notion linguistique de lemme, obtenue par un calcul dans lequel elle n'intervient pas. Elles ne sauraient donner des résultats satisfaisants dans tous les cas. Les « jokers » tel que le symbole * dans les requêtes ci-dessus sont désignés par le terme d'« expressions rationnelles ».

Même lorsque la notion de lemme existe dans le langage de requêtes, des séquences ne comportant pas le lemme recherché peuvent être présentées à l'utilisateur, en raison d'ambiguïtés lexicales. Ainsi, la requête *<aller> à la plage* pourra détecter *le trajet des Allées à la plage*, car pour un logiciel, aucun indice matériel à la fois immédiat et décisif ne distingue *Allées* d'un participe passé du verbe *aller*.

Outre la notion de lemme, d'autres notions linguistiques, telles que les catégories grammaticales, sont manipulables par les langages de requêtes de certains concordanciers, ce qui permet par exemple de produire la concordance de la fig. 2.

Dans la prochaine section, nous examinons par quelles caractéristiques techniques les concordanciers peuvent produire des concordances lemmatisées et plus généralement traiter des requêtes qui font intervenir des notions linguistiques.

Citons cependant trois autres critères importants pour juger de la qualité d'un concordancier.

- Les concordanciers n'ont pas tous les mêmes capacités à **traiter un texte nouveau**. Certains ne peuvent être utilisés que sur des textes dans un format donné. Les différents formats en usage sont liés à l'enrichissement typographique éventuel ainsi qu'au codage des caractères, et notamment des caractères accentués. Il existe des convertisseurs qui font passer les textes d'un format à un autre, quitte à mettre à l'épreuve la débrouillardise de l'utilisateur. On peut donc trouver sur la Toile un texte libre de droits et en produire des concordances quelques minutes plus tard. Il en va autrement lorsque le concordancier ne peut fonctionner que sur des corpus prétraités, par exemple sur des corpus préalablement soumis à une opération d'étiquetage lexical et de lemmatisation, avec révision manuelle. Un tel prétraitement est appelé annotation de corpus, car il consiste à annoter le texte d'informations linguistiques codées. Il s'agit d'un projet de recherche qui peut s'étaler sur plusieurs années car le temps nécessaire à la révision manuelle est proportionnel à la taille du corpus. Les informations linguistiques introduites lors de l'annotation peuvent être exploitées dans les requêtes. En revanche, un tel système ne peut traiter un texte nouveau. Il en est de même des concordanciers qui sont restreints à un corpus spécifique, comme Frantext (É. MARTIN, 1993).

- Les concordanciers offrent plus ou moins de **fonctionnalités statistiques** et comptent par exemple le nombre d'occurrences de chaque mot simple, la répartition d'un mot dans les différentes parties du corpus, les données statistiques sur les collocations...

- Ils diffèrent enfin par leur **interface de présentation** des concordances, qui peut être orientée vers des objectifs particuliers. B. PINCEMIN *et al.* (2006) présentent une recherche spécifique sur ce point.

3. Concordanciers avec dictionnaire

Lorsqu'un concordancier peut traiter des requêtes qui font intervenir des notions linguistiques, et notamment produire des concordances lemmatisées, cela lui permet de mieux cibler ce que l'utilisateur recherche, et donc d'atteindre un meilleur rappel et une meilleure précision. Cela suppose que la notion de lemme a une existence pour le système. Quelles caractéristiques techniques cela implique-t-il ? Quels types de concordanciers possèdent cette fonctionnalité ?

Le premier cas de figure est celui où le corpus est lemmatisé, c'est-à-dire où le lemme de chaque mot du corpus a été précisé, et vérifié manuellement. C'est le cas du Corpus arboré du français (A. ABEILLÉ *et al.*, 2003), qui fait un million de mots. Il existe très peu de corpus lemmatisés du français moderne, comme de la plupart des langues vivantes, et la plupart d'entre eux sont d'une taille trop limitée pour permettre des recherches intéressantes. En effet, la révision manuelle de la lemmatisation d'un grand corpus nécessite beaucoup de temps, de compétence et de travail. Cette première solution ne permet pas non plus de produire à volonté des concordances de textes trouvés sur la Toile. En revanche, de nombreux projets sur des langues anciennes ont opté pour la solution d'un corpus lemmatisé.

Dans la plupart des langues, une seule autre solution permet la génération de concordances lemmatisées : l'utilisation d'un concordancier avec dictionnaire. En effet, si le texte n'est pas lemmatisé, les formes qui figurent dans les textes sont bien des formes fléchies ; et si la requête comporte un mot qui doit être traité comme un lemme, la fonction de recherche du concordancier doit le mettre en relation avec ses formes fléchies, pour pouvoir en détecter les occurrences dans les textes. Cette mise en relation peut se faire de bien des façons, mais elle nécessite toujours un dictionnaire, en raison, par exemple, des nombreuses conjugaisons irrégulières du français.

Plusieurs concordanciers assez connus comportent un dictionnaire. Citons-en trois qui utilisent des stratégies distinctes. Stella (P. BERNARD *et al.*, 2002), le logiciel d'interrogation de Frantext, utilise son dictionnaire pour fléchir les lemmes de la requête et rechercher dans le corpus les formes fléchies obtenues. Cordial Analyseur[2] utilise le sien pour lemmatiser les textes, mais il s'agit ici d'une lemmatisation automatique, sans correction manuelle des erreurs. Unitex[3] (S. PAUMIER, 2002) fait de même, mais en conservant tous les lemmes possibles lorsqu'un mot peut en avoir plusieurs, comme *influent*, qui est une forme du verbe *influer* ou de l'adjectif *influent*. Dans les trois cas, l'utilisateur peut obtenir une concordance telle que celle de la fig. 3.

[2] Cordial Analyseur est un produit commercial de la société Synapse développement. Ses auteurs ne communiquent pas sur son fonctionnement dans des publications scientifiques.

[3] Unitex est un système de traitement de corpus disponible gratuitement et dont la création a fait suite à une innovation algorithmique qui a rendu plus rapide la recherche des motifs linguistiques dans les textes (S. PAUMIER, 2003).

au en ivoire représentant un adolescent jouant aux dés, un saint Joseph en ivoire autref
promettait le club depuis longtemps et jouait au football avec lui chaque dimanche matin
e qui mène à l'Allemagne, les grévistes jouent au football en attendant les premiers cars
toutes cymbales sonnantes. La France jouait à la France. Et, comme s'il n'avait pas su
à donner de lucratives conférences ou à jouer au golf, activités dont l'altruisme ne saut
s de mélodies chinoises et occidentales jouées à la guitare. Pas un slogan aux murs. Le p
administration ; Claire, qui aimerait jouer à la mère incestueuse ; Esther et son insup
devrait donc résister à la tentation de jouer au petit pion, au cuistre traquant les asse

Figure 3. Extrait d'une concordance de séquences *jouer (au + aux + à la + à l')*.

La présence d'un dictionnaire est ainsi une caractéristique essentielle d'un concordancier. La qualité du dictionnaire en question joue également. Sa couverture lexicale et l'exactitude des informations qu'il renferme ont des conséquences directes sur la fiabilité des concordances lemmatisées. Par ailleurs, si le dictionnaire contient des informations syntaxiques, sémantiques ou de domaine, cela ouvre autant de potentialités d'expression de requêtes et augmente l'intérêt des concordances. Ainsi, la concordance de la fig. 4 a été produite à partir d'une requête de la forme *<rapport> de (Dét) <N+hum>*, où le code *+hum* représente le trait sémantique des noms dénotant des personnes, et avec un dictionnaire dans lequel ce trait est marqué. La même requête, sans le trait sémantique, trouve bien d'autres occurrences, dont certaines avec d'autres sens, par exemple *rapports de séduction*.

L'AKP indique qu'un exemplaire du rapport de Mr François Asselineau - qui était en pri
t, tous les neuf ans, la jauge du rapport des forces politiques dans un département, tel
, et resteront fragiles. Selon le rapport du Conseil national des impôts de 1987, 55 % d
ques de l'ensemble du monde ". Le rapport des experts ne souffle mot d'éventuels morts a
de l'autre. On peut lire dans le rapport de Mr Hubert Prévot une série de constatations
u RPCR, s'est bornée à évoquer le rapport de Mr Asselineau sans parler de son contenu et
(le Monde du 6 juillet 1988). Le rapport de la MODAC, commandé au printemps dernier par

Figure 4. Concordance avec trait sémantique.

Pourtant, les concordanciers les plus connus et les plus utilisés n'ont pas de dictionnaire et ne peuvent pas produire de concordances lemmatisées. Ainsi, l'article « concordancier » dans Wikipédia, consulté en octobre 2008, ne cite que des concordanciers sans dictionnaire. Le concordancier le plus utilisé dans les pays de langue anglaise, Wordsmith, n'a pas de dictionnaire non plus. La plupart des utilisateurs de concordanciers ignorent probablement ce que leur apporterait l'utilisation d'un logiciel avec dictionnaire, alors qu'il en existe depuis le début des années 1990, et que certains sont à la fois excellents et disponibles gratuitement. Les utilisateurs se comportent donc comme des consommateurs étonnamment passifs et ne font guère preuve d'esprit critique vis-à-vis de leurs outils[4]. Cette méconnaissance et cette passivité sont paradoxales.

Il est certes facile d'expliquer l'offre abondante de concordanciers sans dictionnaire : ils sont conçus et mis au point par des développeurs de logiciels. Les dictionnaires morpho-syntaxiques et la notion de lemme sont éloignés de l'univers de l'informatique. Les concepteurs de ces outils sont certainement peu attirés par une solution qui leur fait manipuler des notions linguistiques et nécessite des données complexes (les dictionnaires). Ils leur préfèrent la solution alternative des expressions rationnelles (cf. ci-dessus, section 2), qui consiste en un calcul purement informatique, même si cette solution est une approximation assez grossière. Par ailleurs, la promotion des outils réalisés est confiée à des professionnels du marketing, à qui les notions techniques propres au domaine ne sont pas plus familières. Ce sont là des effets du cloisonnement des disciplines et des domaines de compétences, qu'il est banal de constater.

[4] Citons comme contre-exemple l'étude de N. GASIGLIA (2004) qui porte sur Cordial, sur Unitex et sur une utilisation combinée de ces deux outils.

Le paradoxe réside plutôt dans l'attitude peu avisée des utilisateurs, qui sont eux aussi, pour la plupart, des professionnels. Sans entrer dans la sociologie de la communauté scientifique, nous nous contenterons d'avancer deux ébauches d'explications. D'une part, les professions concernées connaissent peut-être une technophobie significative, qui découragerait ceux qui en sont victimes de se pencher sur le fonctionnement des outils. D'autre part, l'utilisation d'un logiciel avec dictionnaire est parfois rejeté comme antinomique avec le recours aux corpus, ceux-ci étant vus comme un gage de rigueur dans l'observation des faits, alors que les dictionnaires seraient perçus, au contraire, comme susceptibles de comporter des erreurs.

4. La flexion dans les dictionnaires électroniques

Les concordanciers avec dictionnaire offrent plus de possibilités, mais le prix à payer est... la nécessité, pour chaque langue concernée, de construire ou de maintenir à jour le dictionnaire. Ces deux tâches n'incombent généralement pas à l'utilisateur final. La recherche publique a financé la construction de dictionnaires électroniques de bonne qualité, dont certains sont maintenant gratuitement disponibles (B. COURTOIS, M. SILBERZTEIN, 1990). Leur mise à jour régulière est une tâche d'intérêt général qui doit être confiée à des lexicologues. Cependant, l'utilisateur peut être amené à gérer un dictionnaire pour un objectif spécifique, par exemple lié à un domaine de spécialité. Dans cette partie, nous examinons comment effectuer cette maintenance. Le lexicologue a deux difficultés principales à résoudre.

La première consiste à détecter et sélectionner les nouveaux mots. Sur ce point, nous renvoyons le lecteur à C. FAIRON et S. PAUMIER (2006). Remarquons uniquement que l'introduction des nouveaux mots dans un dictionnaire électronique nécessite pour chaque entrée une validation par un spécialiste, et insistons sur le fait que cette validation n'est pas entièrement automatisable[5]. Il est donc plus efficace et plus sûr de l'effectuer sur les lemmes que sur les formes fléchies, beaucoup plus nombreuses.

La deuxième difficulté concerne la flexion automatique. Si l'introduction de nouveaux mots et la correction des erreurs se fait au niveau des lemmes, les formes apparaissant dans les textes, elles, sont bien des formes fléchies. Une opération de flexion automatique est donc utilisée pour passer des lemmes aux formes fléchies. Revenons par exemple sur Stella, Cordial et Unitex, les trois concordanciers avec dictionnaire évoqués dans la section précédente. Si la fonction de recherche du concordancier fléchit les lemmes lors du traitement d'une requête, la flexion automatique est nécessaire au fonctionnement du système. Si, au contraire, la fonction de recherche fait appel à un dictionnaire de formes fléchies, ce dernier doit être mis à jour après l'introduction de nouveaux lemmes, et c'est à ce niveau qu'intervient la flexion automatique.

La flexion automatique met en jeu des données linguistiques. Par exemple, la conjugaison en français nécessite des tableaux de suffixes[6] de conjugaison ou des données équivalentes. En allemand et dans les langues slaves, il en est de même pour les déclinaisons. La fig. 5 montre un tableau de suffixes pour la flexion des adjectifs tels que *veuf*.

[5] Le rôle de l'informatique est certes d'automatiser certaines tâches, mais en l'occurrence, et pour des dictionnaires électroniques d'une couverture étendue, comme il en existe pour le français, des décisions d'inclusion de mots nouveaux prises de façon entièrement automatique auraient manifestement un taux d'erreur excessif.

[6] Nous prenons ici le mot suffixe non pas dans son sens linguistique, mais dans un sens algébrique. La délimitation de ces suffixes ne fait pas intervenir un découpage en morphèmes, mais vise seulement à produire les formes voulues par substitution de chaînes de caractères.

masc. sg.	masc. pl.	fém. sg.	fém. sg.
-f	*-fs*	*-ve*	*-ves*

Figure 5. Tableau de suffixes pour la flexion des adjectifs tels que *veuf*.

Ces données linguistiques (ou « ressources linguistiques », dans le jargon du traitement des langues) font elles-mêmes parfois l'objet d'une mise à jour lors de l'introduction de nouveaux lemmes dans un dictionnaire électronique, car leur incorporation dans le lexique de la langue provoque parfois la création d'une variante d'un mode de flexion existant, par exemple dans le cas d'un emprunt étranger qui n'entre bien dans aucun des paradigmes existants. Lorsqu'un nouveau lemme entre dans un des paradigmes existants, ce qui est le cas le plus courant lorsque le dictionnaire a déjà une couverture étendue, il faut spécifier lequel, en lui associant dans le dictionnaire un des paradigmes, par l'intermédiaire d'un identifiant :

veuf,A38

Les logiciels peuvent ensuite procéder à la flexion de façon satisfaisante.

Nous venons de cerner une tâche principale, qui consiste à choisir un paradigme qui convient à un nouveau mot, et une tâche occasionnelle, qui est de mettre à jour les ressources de flexion qui représentent formellement ces paradigmes.

En français et dans la plupart des langues fléchies, ces ressources de flexion prennent des formes différentes suivant qu'elles concernent des mots simples, comme *registre*, ou des expressions multi-mots, comme *main courante*. En effet, la flexion des expressions multi-mots consiste le plus souvent à combiner des formes fléchies des mots simples qui composent le lemme de l'expression. Elle repose donc sur un système de flexion des mots simples. Sur ce sujet, assez technique, nous renvoyons le lecteur à A. SAVARY (2005) pour une présentation de son système Multiflex, et à A. SAVARY (2008) pour un passage en revue critique des systèmes existants.

Pour les mots simples, la plupart des systèmes de flexion automatique utilisent des tableaux de suffixes tels que celui de la fig. 5, mais plus complexes en général. En effet, la fig. 5 se limite à deux traits flexionnels : le genre et le nombre. De nombreux paradigmes flexionnels en mettent en jeu trois, quatre, cinq ou plus, formant des dizaines de combinaisons. De plus, la flexion affecte parfois le radical lui-même, que ce soit dans les langues à suffixes (*tenir/tient*) ou dans les langues sémitiques. Il faut donc soit manipuler des tableaux de formes et de dimensions variées, soit adopter un format plus satisfaisant pour les ressources de flexion.

Cela nous amène à nous demander quels sont les critères de choix d'un format de ressources de flexion automatique. Pour cela, nous devons prendre en compte en priorité les opérations manuelles qui sont effectuées sur ces ressources et qu'il n'est pas possible d'automatiser complètement. Ce sont en effet les opérations qui nécessitent le plus de compétence et de travail d'analyse linguistique de la part des lexicologues à qui elles sont confiées, et donc celles pour lesquelles l'inconfort de l'opérateur est le plus susceptible de provoquer des erreurs. Nous avons déjà évoqué ces tâches manuelles plus haut : il s'agit

- d'identifier, parmi des ressources existantes représentant des paradigmes, celle qui convient à la flexion d'un nouveau lemme, s'il en existe ;

- à défaut, de construire une nouvelle ressource représentant un nouveau paradigme.

Le critère de qualité essentiel est à notre avis la lisibilité visuelle, comme pour tout format de ressources linguistiques sur lequel on effectue des opérations manuelles. La ressource doit être suffisamment lisible pour que l'opérateur interprète facilement et correctement ce qu'elle représente. L'expérience prouve que l'on supprime ainsi la principale source d'erreurs. En particulier, il est inévitable qu'il y ait des conventions d'interprétation, mais il est préférable qu'elles soient accessibles à des utilisateurs qui ne connaissent aucun langage de programmation, car il n'y a pas de raison particulière pour qu'un bon lexicologue se trouve être également développeur de logiciel.

Un second critère de qualité pour le format des ressources est qu'il soit applicable à toutes les catégories grammaticales et à des langues aussi nombreuses que possible. En effet, s'habituer à un format et aux conventions qui lui sont attachées est un investissement ; utiliser le même format pour d'autres catégories grammaticales ou une autre langue, c'est rentabiliser et approfondir cet investissement.

Vis-à-vis de ces deux critères, les tableaux de suffixes nous semblent inférieurs à un autre format de ressources, celui des « transducteurs de flexion » introduits par M. SILBERZTEIN (1998 : 187-189) et éditables avec Unitex (fig. 6).

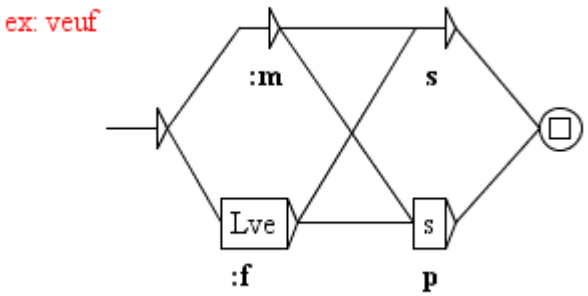

Figure 6. Transducteur de flexion pour les adjectifs tels que *veuf*.

Le transducteur de flexion de la fig. 6 comporte les mêmes informations que le tableau de la fig. 5, mais les présente d'une autre façon. Chaque forme fléchie correspond à un chemin qui va du nœud initial, qui est la flèche de gauche, jusqu'au nœud final, qui est le cercle de droite. Le chemin inférieur, par exemple, passe par un nœud qui contient la chaîne de caractères *Lve* (ce qui représente la suppression de la lettre finale du lemme puis l'ajout de *ve*), puis par un autre qui contient la lettre *s*. Les opérations représentées par ce chemin peuvent être récapitulées sous la forme *Lves*, ce qui signifie : supprimer la lettre finale du lemme, puis ajouter *ves*. Les codes affichés sous ces deux nœuds, **:f** pour féminin et **p** pour pluriel, peuvent de même être récapitulés sous la forme **:fp**, code du féminin pluriel. Les opérations de substitution de suffixes, affichées dans le contenu des nœuds, sont ainsi mises en relation avec les codes flexionnels affichés sous les nœuds. Il en est de même des autres chemins obtenus en suivant les transitions de gauche à droite de toutes les façons possibles. En énumérant tous les chemins, on obtient bien les quatre formes fléchies voulues et leurs codes flexionnels :

veuf	A:ms
veufs	A:mp
veuve	A:fs
veuves	A:fp

Pour le lecteur, ces conventions de représentation rendent peut-être le transducteur moins lisible que le tableau de la fig. 5. Cependant, elles ont l'avantage d'être directement applicables à des paradigmes beaucoup plus complexes. Prenons une conjugaison en français, représentée dans la fig. 7 avec quatre traits flexionnels : temps-mode, personne, genre et nombre.

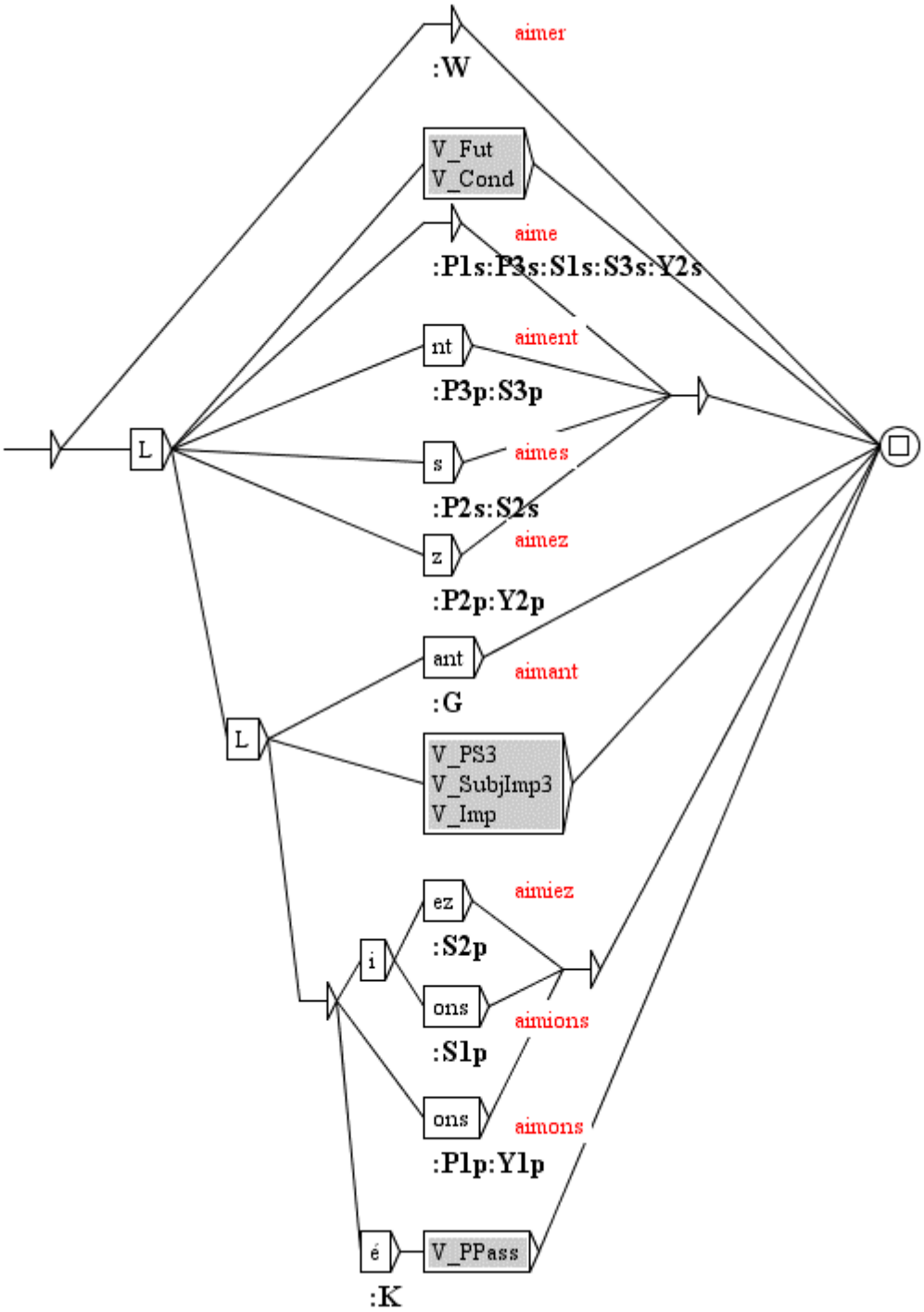

Figure 7. Transducteur de flexion pour les verbes tels qu'*aimer*.

Le même paradigme représenté dans le format de la fig. 5 prendrait la forme d'un tableau à quatre dimensions, qu'il serait complexe d'afficher à l'écran ou d'imprimer. Dans la fig. 7, les formes du verbe *aimer* apparaissant en dehors des nœuds sont de simples aides à la lecture, ignorées par le logiciel de flexion automatique. Les codes du type **:P3p:S3p** correspondent au cas où une forme a deux interprétations qui correspondent à des codes morpho-syntaxiques différents : ici, **P3p** pour le présent de l'indicatif et **S3p** pour le présent du subjonctif. Les nœuds à contenu grisé invoquent des sous-transducteurs qui contiennent des ensembles de suffixes communs à plusieurs conjugaisons, par exemple les suffixes du futur et du conditionnel.

Ce format, né à la suite d'une quinzaine d'années d'expérience de construction et de maintenance manuelles de dictionnaires électroniques au LADL[7] sous la direction de Maurice Gross, a fait ses preuves sur les langues à flexion par suffixes. Il est extensible aux langues sémitiques et aux langues agglutinantes, en ce sens que des transducteurs de flexion et des grammaires de morphèmes respectant des conventions très voisines peuvent être utilisés pour ces langues.

Conclusion

Nous avons brièvement présenté une méthode éprouvée de flexion automatique de dictionnaires, et nous en avons montré l'intérêt pour la maintenance d'un concordancier capable d'engendrer des concordances lemmatisées. Un tel concordancier, à son tour, est utile, entre autres, aux lexicologues. Enfin, l'ensemble de ce schéma est applicable à de nombreuses langues.

Ce réseau de méthodes et de flux d'informations est un exemple de coopération entre linguistique et informatique où chaque opération est à sa place. Les algorithmes informatiques sont utilisés pour offrir au lexicologue un environnement professionnel qui lui permet, dans un cas, de procéder à de la description lexicologique, et dans l'autre, d'engendrer rapidement des concordances. Les ressources linguistiques sont exploitées pour élever le rappel et la précision des concordances, et satisfaire ainsi leurs utilisateurs.

Dans les pratiques effectives des chercheurs, l'articulation entre linguistique et informatique n'est pas toujours aussi rationnelle et efficace que dans cet exemple. Les algorithmes informatiques servent trop souvent à obtenir par le calcul des résultats approximatifs, alors même qu'il est possible d'obtenir des résultats plus exacts en utilisant des dictionnaires gratuitement disponibles. Inversement, les lexicologues se contentent trop souvent d'outils informatiques pour lesquels le seul objet de base est le mot simple fléchi, à l'exclusion du lemme et du mot composé, pourtant tout aussi pertinents pour eux, sinon plus.

Nous, spécialistes du traitement des langues, avons un rôle à jouer pour lutter contre ces anomalies dans l'articulation entre linguistique et informatique. Notre connaissance interdisciplinaire nous confère la responsabilité de mieux définir les objectifs de l'informatique pour les linguistes : il s'agit de cibler ces objectifs vers l'automatisation de tâches qu'il est réellement utile et possible d'automatiser. Par ailleurs, nous devons fournir à ceux qui se définissent sans ambiguïté comme des linguistes les moyens de faire des choix éclairés en matière d'outils informatiques, en attirant leur attention sur certains critères de choix dont l'importance n'est pas facile à déceler.

C'est ce que nous avons tenté de faire ici.

Références bibliographiques

ABEILLÉ, Anne, Lionel CLÉMENT, François TOUSSENEL (2003) : « Building a treebank for French », in ABEILLÉ, Anne (ed), *Treebanks*, pp. 165-187. Dordrecht, Kluwer.

BERNARD, Pascale, Josette LECOMTE, Jacques DENDIEN, Jean-Marie PIERREL (2002) : « Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis : Frantext, TLFi, and the software Stella », in *Proceedings of the Conference on Language Resources and Evaluation* (LREC), pp. 1090-1096, Las Palmas, Espagne.

[7] Laboratoire d'automatique documentaire et linguistique, Université Paris 7, 1968-2000.

CABALLERO, María del Rosario (1999) : « Using a Concordancer in Literary Studies », *The European English Messenger*, VIII(2), pp. 59-62. http://www.edict.com.hk/Concordance/

COURTOIS, Blandine, Max SILBERZTEIN, éds. (1990) : *Langue française*, 87. Paris, Larousse.

FAIRON, Cédrick, Sébastien PAUMIER (2006) : « A framework for real time dictionary updating », in *Proceedings of the Language Resources and Evaluation Conference* (LREC), Genoa.

GARRIGUES, Mylène (1998-1999) : « Nouvelles concordances pour l'enseignement des langues », *Lingvisticae Investigationes*, 22, pp. 59-69. Amsterdam/Philadelphie, John Benjamins.

GASIGLIA, Nathalie (2004) : « Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus », *Revue française de linguistique appliquée*, IX, pp. 45-62.

GOUADEC, Daniel (1997) : *Terminologie et phraséologie pour traduire. Le concordancier du traducteur*. Terminoguide nº 3 - Traduguide nº 3. Paris, La Maison du Dictionnaire.

GROSS, Maurice (1989) : « La construction de dictionnaires électroniques », *Annales des Télécommunications,* 44(1-2), pp. 4-19.

GROSS, Maurice (2000) : « A bootstrap method for constructing local grammars », in Neda BOKAN (ed.), *Contemporary Mathematics. Proceedings of the Symposium*, pp. 229-250. 18-20 December 1998, University of Belgrade.

JACQUET-PFAU, Christine (1994) : « L'intérêt des logiciels de concordances pour la traduction », *Langages*, 116, pp. 82-86, Paris, Larousse.

MAGRI-MOURGUES, Véronique, éd. (2006) : *Corpus*, 5, Corpus et stylistique. http://corpus.revues.org/index.html

MARTIN, Éveline, coord. (1993) : *Frantext. Autour d'une base de données textuelles. Témoignages d'utilisateurs et voies nouvelles*. Paris, Didier-Érudition.

PAUMIER, Sébastien (2002) : *Unitex. Manuel d'utilisation.* Édition en anglais révisée pour la version 2.0 (2008), http://igm.univ-mlv.fr/~unitex/UnitexManual.pdf. Université Paris-Est Marne-la-Vallée.

PAUMIER, Sébastien (2003) : « A Time-Efficient Token Representation for Parsers », *Proceedings of the EACL Workshop on Finite-State Methods in Natural Language Processing*, Budapest, pp. 83–90.

PINCEMIN, Bénédicte, Fabrice ISSAC, Marc CHANOVE, Michel MATHIEU-COLAS (2006) : « Concordanciers : Thème et variations », *Lexicometrica*, numéro spécial, pp. 769-780. Actes des Journées d'analyse statistiques des données textuelles (JADT), http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/tocJADT2006.htm.

RENOUF, Antoinette, Andrew KEHOE, Jay BANERJEE (2007) : « WebCorp: an integrated system for web text search », in Marianne HUNDT, Nadja NESSELHAUF, Carolin BIEWER (eds.), *Corpus Linguistics and the Web*, pp. 47-67. Language and Computers, 59. Amsterdam: Rodopi.

SAVARY, Agata (2005) : « A formalism for the computational morphology of multi-word units », *Archives of Control Sciences* 15(3), pp. 437–449. Gliwice (Poland), Silesian University of Technology.

SAVARY, Agata (2008) : « Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches », in *Linguistic Issues in Language Technology.* http://csli-publications.stanford.edu/LiLT/

SILBERZTEIN, Max (1993) : *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Paris, Masson.

SILBERZTEIN, Max (1998) : « INTEX: An integrated FST toolbox », in Derick WOOD, Sheng YU (éd.), *Automata Implementation*, p. 185-197, Lecture Notes in Computer Science, vol. 1436. Second International Workshop on Implementing Automata (1997), Berlin/Heidelberg: Springer.

SINCLAIR, John, ed. (1987) : *Looking Up: an account of the COBUILD Project in Lexical Computing*. London, Collins.

TOGNINI-BONELLI, Elena (2001) : *Corpus Linguistics at Work*. Amsterdam/Philadelphia, John Benjamins.

TRIBBLE, Chris, Glyn JONES (1997) : *Concordances In The Classroom. A resource book for teachers.* Houston: Athelstan.

ZINGLÉ Henri (1994) : « The ZStation workbench and the modelling of linguistic knowledge », in Carlos MARTIN-VIDE (ed.), *Current Issues in Mathematical Linguistics*, pp. 423-432. Selected Papers from the 1st International Conference on Mathematical Linguistics, Tarragona, Spain, 30-31 March 1993. North Holland Linguistic Series. Amsterdam, Elsevier-North Holland.